AMENDMENT TO THE SPECIFICATION

All references to the application will refer to the published version at U.S. Pre-Grant Publication No. 2005/0216253.

Please amend paragraph [0035] as follows:

> Each of the words in word pair set 222 is operated on, if necessary, by tokenizer 224 in order to segment the word into component characters, or sequences of frequently co-occurring characters, for example, the English letter sequence "qu", in each respective word, where "characters" as used herein is to include all component parts of words used in any language, e.g. English, Japanese, Chinese, Arabic, etc. A clustering system 225 can optionally operate on the word pair sets 222 to provide hierarchical clustering of characters. This benefits the system by boosting probabilities of alignments when characters have similar contextual associations. An exemplary clustering algorithm (JCLUSTER) is available~~ at~~ ~~http://www.research.microsoft.com/research/downloads/~~, although many other clustering algorithms can be used. In any case, the word pair sets 222 are provided to character alignment system 210.

Please amend paragraph [0038] as follows:

> This offers several advantages. For example, it permits the system to be used between language pairs for which phonological data may not exist, or when phonological information is not

available, for example, Arabic or Chinese names when encountered in Japanese, but which need to be identified in English. Furthermore, because alignment system 210 uses standard machine translation techniques, the direction of mapping is completely and immediately reversible~~reversable~~, allowing the relationship between the languages to be reversed with the same training data. A further advantage of the machine translation modeling over simple character correspondence of word pairs or phonological models is the ability to map characters to null characters; among other things, this permits the system to be relatively robust when confronted with noisy morphological variation between the two languages as might be encountered when data is extracted from parallel texts. For example, given a Japanese katakana form "" that can be directly transliterated under one conventional transliteration scheme as "ma-ne-e-ji", the alignment system 210 can learn that these characters map to the English word "managed" in certain contexts, e.g., English "managed code", despite the additional "-ed" which lacks any counterpart in the Japanese; likewise, the system is able to learn the relevant alignments between the characters in the Japanese word "", directly transliterated under one conventional transliteration scheme as "i-n-su-to-o-ru" and English "installation". FIG. 4A pictorially illustrates the alignments for this latter word pair, learned under one embodiment of the system. In this example, several characters in the English word, namely those in the final character sequence "a-t-i-o-n-$", are aligned to the Japanese end-token "$", allowing this English sequence to be potentially available to a cognate word identification system such as that in 211, albeit with a lower likelihood. This robustness, inherited from statistical machine translation, permits alignment system 210 to learn contextual mappings directly from ordinary parallel text data, something that phonological systems cannot do.

Please amend paragraph [0049] as follows:

In yet another application, the system might be used as a component of an automated glossing application to assist reading of a foreign language word, by allowing for example a user to place a computer cursor over a word on a web page or other document to pop up a translation. In this application, the system would supplement existing bilingual lexical lookup or machine translation by providing the additional functionality of identifying candidate proper names and other terms that are not in a dictionary.